面向中文图书评论的情感词典构建方法研究

郭顺利 张向先
(吉林大学管理学院 长春 130022)

摘要：【目的】探讨中文图书评论情感词典构建方法，以便进行用户图书评论的情感分析。【方法】参照相关研究将用户情感分为 7 类，对采集到的语料库进行分词，结合基础情感词典得到中文图书评论的情感词集，选取各类情感种子词；利用改进的 SO-PMI 算法和同义词词林扩展方法判别词语的情感类别；以实际的图书评论作为语料进行实验验证。【结果】提出一种中文图书评论的情感词典构建方法，其平均准确率、平均召回率及 F1 的均值分别为 0.90、0.83 和 0.85。【局限】语料库小，样本范围具有一定的局限性。【结论】实验结果表明本文方法具有较高的有效性和可靠性，能够有效地进行用户图书评论的情感分析。

关键词：中文图书评论 情感词典 种子词 情感分类 SO-PMI 算法
分类号：G353

1 引 言

随着互联网的发展，越来越多的用户通过网络社交平台对日常新闻事件、产品、政策制度等发表个人观点和意见，从而形成用户评论。用户评论中含有大量的情感词语，能够体现出用户的个人情感。在商务领域，用户评论是指用户购买或体验某一产品后对产品和服务做出的评价，能够体现出用户对产品的个人情感信息，通常用于商业反馈。用户图书评论是指用户对于某一本图书发表的评论或介绍书籍的文本，是以"书"为对象，实事求是、有识见地分析书籍的形式和内容，探求创作的思想性、学术性、知识性和艺术性，从而在作者、读者和出版商之间构建信息交流的渠道[1]，即用户阅读某本书后对于书中内容的评价以及个人情感观点的表达，能够体现出用户对于图书的情感信息。利用用户图书评论进行情感分析能够更好地挖掘用户行为，为图书的发行出版以及其他用户的阅读选择提供建议。如何快速准确地对大量的用户图书评论进行情感分析成为重要的研究课题。对于用户图书评论进行情感分析需要用到情感词典，目前国内还没有一部完善的大规模中文图书评论情感词典。中文图书评论情感词典是进行中文图书评论情感分析的前提，如何从大量的用户图书评论获取情感词汇，自动构建中文图书评论的情感词典，已成为中文图书评论情感分析研究亟需解决的问题。

2 国内外研究现状

在情感词典构建研究中，国外的研究人员主要基于 WordNet 词典进行英文情感词典的构建研究[2]，Turney通过改进PMI-IR算法进行无监督的情感分析并取得较好的效果[3]。Subasic等手工构建一个基于情感类别相关的词典，词典中标明了词的强度(表达情感的力度)和向心度(与类别的相关程度)[4]。目前国内中文情感词典的构建研究工作也取得了部分成果。借助HowNet、《同义词词林》等词典，在HowNet的基础上构建特定情感词典的研究也有很多。例如柳位平等在中文词语相似度计算方法的基础上，提出一种中文情感词语的情感权值的计算方法，并以情感词语集为基

通讯作者：郭顺利，ORCID：0000-0003-3186-2677，E-mail：ssguoshunli@sina.com。

准，构建中文基础情感词典[5]。国内对于微博情感词典的构建研究相对较多，不同的学者利用不同的方法从不同的角度进行微博情感词典的构建。李钰整合基础情感词典、虚词词典、表情符号情感词典和网络用语情感词典得到微博情感词典[6]。桂斌等基于微博表情符号，提出一种自动构建情感词典的方法[7]。也有相关的学者对于不用的领域，构建相关的领域情感词典，如周咏梅等借鉴图排序模型的原理，提出一种新闻评论情感词典构建方法[8]。蒋盛益等利用改进后的Hevner情感环模型为基础，借助HowNet所提供的语义资源和从网络爬取的歌词文本语料库，构建了一部音乐领域中文情感词典[9]。还有其他领域的相关情感词典，例如酒店评论情感词典[10]、微博产品评论情感词典[11]、电影评论情感词典[12]等。

笔者经过调研发现面向中文图书评论领域情感研究很少，其相关的情感词典研究几乎空白。中文图书评论不同于其他领域的用户评论，它使用的很多词语具有较强的文学性和学术性，有些情感词语在其他领域评论中很少使用，具有一定的专业特色。例如"讶异"、"妙笔生花"等词。同时中文图书评论具有固定的书写格式，拥有一定的规范性。因此其他领域的情感研究难以有效地应用于中文图书评论的情感分析研究，使其具有一定的研究价值和意义。因此，本文提出一种中文图书评论情感词典的构建方法，并构建一部中文图书评论情感词典，以便于后续的中文图书评论的情感分析。

3 中文图书评论情感词典构建思路概述

本文提出的中文图书评论情感词典构建方法的基本流程如图1所示。

(1) 参照文献[13]中的情感分类方法将中文图书评论的情感分为7大类。

(2) 利用ROST CM6分词工具[14]将采集到的中文图书评论语料库进行分词和词频统计，结合基础情感词典进行比较分析后综合得到中文图书评论情感词集。

(3) 中文图书评论7大类种子情感词的产生。在产生的情感词集基础上查询情感词汇本体中情感词的强度，结合情感词集中情感词词频，利用人工筛选判定的方法，得到中文图书评论7大类情感的种子词。

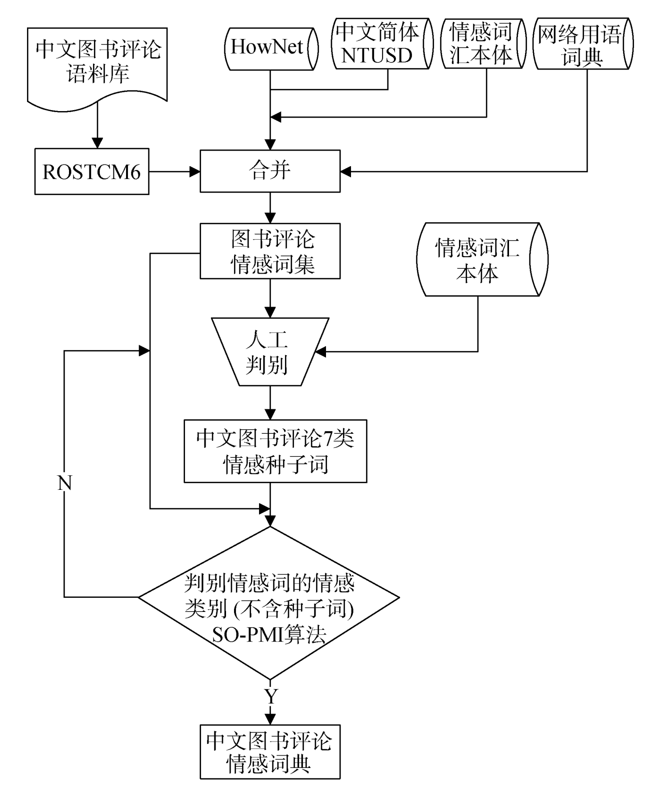


图1 中文图书评论情感词典构建流程

(4) 利用改进的SO-PMI算法对中文图书评论情感词集中的情感词语(除了种子词以外)进行情感归类，通过同义词词林[15]扩展的方法解决数据稀疏问题。得到每个情感词语的关联度，根据关联度的大小将情感词语归到中文图书评论7类情感类别中。

(5) 得到中文图书评论情感词典。

4 基于SO-PMI算法的中文图书评论情感词典构建研究

4.1 中文图书评论情感分类

中文图书评论能够体现出读者的情感，同时不同类型的图书与不同的情感类别相联系，因此有必要将中文图书评论体现出的用户情感分成不同的类别。目前，心理学界对于情感的分类没有公认的标准，研究者将情感分为4、6、8、10乃至20余类不等。本文参照文献[13]中的用户情感分类方法，综合现有的情感词汇资源，将中文图书评论的用户情感分为7类，分别为乐、好、怒、哀、惧、恶、惊。

4.2 中文图书评论情感词集的产生

互联网上用户的图书评论中使用了一些网络新词

和网络用语，现有的情感词典不能够覆盖用户图书评论中所有的情感词。因此本文通过整合现有的情感词典、用户图书评论语料库中的情感词语以及互联网上的网络用语，构建用户图书评论的数据。在中文图书评论情感词集构建过程中，主要使用的情感词典资源有：大连理工大学信息检索研究室的情感词汇本体[13]、HowNet 情感词典[16]和台湾大学的中文情感极性词典 NTUSD[17]、网络情感词词典，其中网络情感词典采用人工收集的方式进行构造。在中文图书评论的情感词集的构建过程中，利用爬虫软件集搜客 GooSeeker[18]从豆瓣网上的豆瓣读书中爬取文学、流行、文化、生活、经管、科技类的图书评论。经过去除中性评论、垃圾评论和表情符号的转换等预处理，共收集到关于 8 500 本图书，总计 255 000 余条的用户图书评论数据。利用 ROST CM6 分词工具进行切词和词频统计，经过人工筛选和判断去掉词频数量较低的词语，形成中文图书评论词集 WordSet1 。将情感词汇本体、HowNet 情感词典和中文情感极性词典 NTUSD、网络情感词词典中的情感词语合并构成基础情感词典词数据 WordSet2，将获得的中文图书评论词集 WordSet1 与基础情感词典词集 WordSet2 中的词语进行比较，取两个词集的交集形成中文图书评论情感词集 WordSet。经过以上的处理后得到的中文图书评论词集 WordSet 含有 881 个情感词语。

4.3 种子情感词的选择

本文拟采用一种改进的 SO-PMI 算法进行词语情感倾向性判断，所以需要进行种子词的选择。这里的种子词是指情感态度非常明显、强烈、具有代表性的词语。在 4.2 节得到的中文图书评论情感词集 WordSet 的基础上，查询这些词语在情感本体中的强度，将强度最大且在语料中出现频率较多的词作为候选种子情感词，例如“团圆”这个词语在情感本体中的情感强度为 9（最强），情感分类小类为快乐，大类为乐，而在中文图书评论语料库中统计频率为 4 782 次，则将它作为乐的候选种子情感词。经过上述的选择处理，得到 7 类情感的种子情感词共计 191 个，形成种子词集 S，如表 1 所示：

表 1 中文图书评论 7 类情感种子词

情感类别	情感种子词
乐(16)	得意洋洋 皆大欢喜 痛快淋漓 大功告成 喜滋滋 过瘾 随心所欲 舒畅 欣喜若狂 令人满意 晋升 得心应手 喜气洋洋 开心 愉快 团圆
好(59)	主力 至上 美好 令人钦佩 令人神往 痛痛快快 鬼斧神工 仁慈 英雄 侠客 鲜花 财宝 义无反顾 完满 倾注 妙笔生花 秀外惠中 平易近人 善良 讴歌 英明 功不可没 珍重 和谐 珍惜 史无前例 歌功颂德 痛快淋漓 力挽狂澜 别具一格 拯救 美妙 珍贵 创新 灿烂 推崇 赞许 英俊 过人 侠义 完备 出类拔萃 文武双全 推荐 推进 开朗 辩护 漂亮 令人信服 倾心 珍藏 倾倒 珍宝 至亲 陶醉 珍品 珍视 珍爱 法宝
怒(17)	咬牙切齿 杀气 勃然大怒 恼羞成怒 精疲力竭 筋疲力尽 投诉 血债 肝火 鬼哭狼嚎 怒气冲冲 创巨痛深 一无所有 怒火 愤怒 气急败坏 怒骂
哀(17)	受害 舍弃 创伤 血泪 亡国 血案 遍体鳞伤 疮痍 自惭形秽 追悔莫及 千头万绪 缅怀 眼睁睁 无影无踪 拒绝 悲惨 拜拜
惧(16)	紧绷绷 草木皆兵 团团转 财政危机 心慌意乱 心急火燎 灾难性 鬼哭神嚎 害臊 顷刻 提心吊胆 受惊 绝望 令人不安 悬崖峭壁 惊魂未定
惊(11)	目瞪口呆 叹为观止 奇怪 奇迹 瞠目结舌 大吃一惊 震撼人心 大惊失色 魂消胆丧 讶异 魂惊胆颤
恶(55)	恶毒 邪恶 恶梦 窒息 凶恶 恶心 受过 勉强 两难 受苦 受挫 受骗 受制 令人发指 令人作呕 罪不容诛 憎恨 肆虐 憎恶 猖狂 铁蹄 推脱 杀伤 杀戮 勾当 完蛋 违心 一无是处 纸醉金迷 昭然若揭 胡作非为 雪上加霜 草菅人命 里通外国 污七八糟 违背 庸俗 帝国主义 害人 蛊惑 罪恶 过激 忘恩负义 孤芳自赏 殴打 大言不惭 血腥 违反 抨击 丑恶 盗窃 模棱两可 觊觎 辩驳 霸占

4.4 基于改进的 SO-PMI 算法的情感词情感类别判断方法

词语情感类别判断的方法主要有基于 HowNet 的语义相似度计算方法以及基于 SO-PMI 的情感词倾向性计算方法[19-20]。因为用户的图书评论中存在大量的网络新词，例如“给力”、“正能量”、“坑爹”等在 HowNet 中找不到义原，从而也就无法根据两个词义原的相似度计算词语的相似度，所以基于 HowNet 的语义相似度计算方法进行中文图书评论中部分词的情感类别判断并不适用。所以本文提出一种改进的基于 SO-PMI 的词语情感类别判别方法，通过互信息计算未知词与各类种子词关联度的方法对未知词的情感类

别进行判断。如公式(1)所示:

$$PMI(word_1, word_2) = \log \frac{P(word_1, word_2)}{P(word_1)P(word_2)} \quad (1)$$

其中，$PMI(word_1, word_2)$ 表示 $word_1$ 与 $word_2$ 的关联度，$P(word_1, word_2)$ 表示 $word_1$ 与 $word_2$ 共现的概率，$P(word_1)$ 表示 $word_1$ 在语料库中出现的概率，$P(word_2)$ 表示 $word_2$ 在语料库中出现的概率。

使用词语出现的次数代替出现的概率，由于两词语共现之间的距离与两词语的关联强度成反比，即两词语离得越近，关联度越大；反之，两词语离得越远，关联度越小。应用在词语的情感倾向性分析中，就是两词语离得越近，情感倾向性相关的可能性越高。两个词语之间的距离用两个词语之间的字符数量表示，把两个词语在同一评论中距离的最小值作为两个词语的共现距离 d，两个词语之间的共现距离 d 计算公式如下:

$$d = \min |d_x - d_y| \quad (2)$$

其中，d 表示两个词语之间的共现距离，d_x 表示在每条评论中从评论开始到两个词语排在前面词语的最后一个字符的字符个数，d_y 表示在每条评论中从评论开始到两个词语排在后面词语的第一个字符的字符个数。

为此本研究将两个词语之间的共现距离 d 加入到互信息计算公式中则可以将公式(1)变换成公式(3):

$$PMI(word_1, word_2) = \log_2\left(\frac{N \times hit(word_1, word_2)}{d \times hit(word_1) \times hit(word_2)}\right) \quad (3)$$

其中，N 表示语料库中所有词语的总次数，hit 表示词语的词频数。$hit(word_1, word_2)$ 表示词语 $word_1$ 和 $word_2$ 在以同一本书的评论为单位中总计共现的次数。d 表示两个词语之间的共现距离。

中文图书评论情感词集 WordSet 去除种子词后形成需要判断的图书评论情感词集为 WordSetX 。未知情感词 $word_x$ 与每一类情感之间的情感词关联度 $SO_PMI(word_x, S_i)$ 是未知的情感词 $word_x$ 和第 i 类情感的种子词集合 S_i $(1 \leqslant i \leqslant 7)$ 中的每个种子词的 PMI 之和，如公式(4)所示:

$$SO_PMI(word_x, S_i) = \sum_{s_a \in S_i} \log_2\left(\frac{N \times hit(word_x, s_a)}{M_i \times d \times hit(word_x) \times hit(s_a)}\right) \quad (4)$$

其中，N 表示语料库中所有词语的总次数，hit 表示词语的词频数。s_a 表示第 i 类情感种子词集合 S_i 中的第 a 个情感词，$hit(word_x, s_a)$ 表示词语 $word_x$ 和 s_a 在以同一本书的评论为共现窗口中共现的次数。$M_i (1 \leqslant i \leqslant 7)$ 表示第 i 类情感种子词集 S_i 中种子词的数量。

由于每类情感的种子词数量不同，可能会出现未知情感词并不与该类最相关，也会因为该类情感情感种子词多造成累加项较多，从而使得总的相关度最高。为了避免出现此类偏差，使用 $M_i (1 \leqslant i \leqslant 7)$ 表示第 i 类情感种子词集 S_i 中种子词的数量，将 M_i 加入到公式消除此类偏差。

由于利用互信息进行关联度的计算会受到语料库规模的影响，同时也因为用户写作习惯的不同，不同的用户在撰写评论的时候，可能会使用不同的词语表达相同的意思。即使同一用户在撰写评论中也常常会使用同义词表达相同的意思。所以仅仅考虑一个词语与种子词的共现信息就可能出现数据稀疏问题。因此本研究在计算候选词情感极性时，通过同义词词林扩展版对候选词进行扩展，从而减少某些词在语料库中出现频率太低所带来的数据稀疏问题。根据相关的实验随着扩展次数的增加会致使原词损失语义，所以本研究将扩展迭代次数设为三次。

利用本文提出的改进 SO_PMI 算法进行词语的情感类别分类的算法如下:

输入：WordSetX，S，同义词词集 SameWord，N，$M_i (1 \leqslant i \leqslant 7)$。

输出：WordSetX 中的未知情感词 $word_x$ 的情感分类。

①从 WordSetX 中取出未知情感词 $word_x$，$1 \leqslant x \leqslant 690$，种子词集中每类情感的种子词集为 S_i $(1 \leqslant i \leqslant 7)$。

②按照公式(4)计算情感词 $word_x$ 分别对于用户图书评论 7 类情感的关联度 $SO_PMI(word_x, S_i)$。

③把步骤②中计算出的 $SO_PMI(word_x, S_i)$ 按照从大到小的顺序进行排列，取最大值作为判断 $word_x$ 归属于第 i 类情感的依据。

④如果步骤③中最大的 $SO_PMI(word_x, S_i)$ 不为 0，则跳转到步骤⑥；如果最大的 $SO_PMI(word_x, S_i)$ 为 0，则运用同义词词集 SameWord 对 $word_x$ 进行同义词扩展，找出 $word_x$ 的同义词集合 WordSameB，计算出同义词的个数 B，则 $1 \leqslant b \leqslant B$ 分别计算 WordSameB 中每一个词语与 7 类情感的归属度 $SO_PMI(WordSame_b, S_i)$。

⑤跳转到步骤③进行判断 $word_x$ 归属于哪类情感，如果最大的 $SO_PMI(WordSame_b, S_i)$ 仍为 0，则利用步骤④中的方法对 $word_x$ 的同义词集合 WordSameB 进行进一步的同义词扩展，扩展次数最多为三次，若得到最大的 $SO_PMI(WordSame_b, S_i)$ 仍为 0，说明词语的情感强度较弱，

直接删除。

⑥算法输出，输出 $word_x$ 的情感归类情况。

⑦算法结束。

5　实验分析

为了验证本研究中文图书评论情感词典构建方法的有效性，从词语情感类别判定准确性和基于构建的中文图书评论情感词典分类性能两个方面进行具体的实验验证。利用 GooSeeker 爬虫[18]从豆瓣网上采集 100 本图书的图书评论，共计有图书评论 15 000 余条，形成语料库。经过预处理和数据清洗选取其中的 5 000 条图书评论进行实验，将这 5 000 条图书评论的 7 类情感分类情况进行人工标注。将利用 4.2 节得到的去掉种子词后的 690 个中文图书评论情感词进行人工标注情感类别。

5.1　判定准确性实验

通过查询情感本体强度表和人工判断，得到690个情感词的人工判别分类情况，利用原有的SO-PMI算法进行判别，利用4.4节改进的SO-PMI算法进行情感词的判断，得到改进SO-PMI算法的分类情况，结果如表2所示：

表 2　7 类情感的情感词数量分布表

情感分类	乐	好	怒	哀	惧	恶	惊
人工判别	58	207	62	67	59	199	38
SO-PMI 算法判别	48	221	51	65	57	218	30
SO-PMI 算法正确判别	38	196	39	47	43	170	20
改进 SO-PMI 算法判别	52	213	58	73	50	211	33
改进 SO-PMI 算法正确判别	49	204	46	62	48	196	26

本组实验采用准确率(P)、召回率(R)、F1值(F1)三个评估指标进行改进的SO-PMI算法与原来的SO-PMI算法之间进行方法的性能比较，经计算结果如表3所示。并利用SPSS 19.0对两种方法的判别情况绘图，如图2所示。

表 3　7 类情感词的 SO-PMI 算法性能评估

指标 \ 分类		乐	好	怒	哀	惧	恶	惊	总体
SO-PMI 算法	P	0.79	0.89	0.76	0.72	0.75	0.78	0.67	0.77
	R	0.66	0.95	0.63	0.70	0.73	0.85	0.53	0.72
	F1	0.72	0.92	0.69	0.71	0.74	0.81	0.59	0.74
改进 SO-PMI 算法	P	0.94	0.96	0.79	0.85	0.96	0.93	0.79	0.89
	R	0.84	0.99	0.74	0.93	0.81	0.98	0.68	0.85
	F1	0.89	0.97	0.76	0.85	0.88	0.95	0.73	0.86

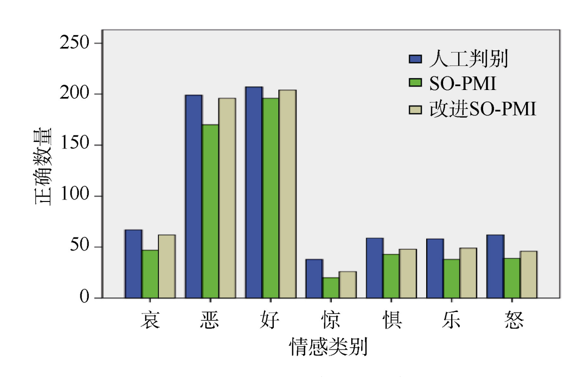


图 2　两种方法判别 7 类情感的情感词分布对比

从表2、表3以及图2可以看出，本文改进的SO-PMI算法进行情感词的情感判别准确率平均值为0.89，召回率平均值为0.85，F1值的平均值为0.86。其准确率、召回率和F1值均比原有的SO-PMI算法高，所以利用改进的SO-PMI算法进行情感词情感判别比原有的SO-PMI算法判别方法效果好，所以总体看来本文中提出的情感词情感类别判断方法具有较高的准确性和可利用性。例如：在进行情感词类别判断时，对于情感词集中“价值”这个词，图书评论语句如下：

> 这本书有一定的价值，非常值得推荐。作者用悲惨的结局。来解释缺失的童年，来深入读者的内心产生共鸣，在阅读的过程我没有流泪，是不是我特别的冷血？但写的比我预期的好。

在对“价值”这个词语进行情感归类时，按照4.4节提出的计算方法，“价值”归于情感类别“好”，“哀”的种子词数量不同于“好”的种子词数量，所以需要考虑到种子词的数量，同时如果不加入两词之间的共现距离d，其计算的关联度中属于“哀”的关联度大于属于“好”，出现一定的判别误差。所以加入共现距离d和种子词的数量 M 能够提高算法的准确性和可利用性。

在进行词语情感类别判断时，利用同义词词林进行扩展能够解决出现的数据稀疏性问题，例如：对于评论“看到一半放弃了，太苦，有些东西也无法认同。”在判断词语“太苦”时，发现评论太短，不存在选取的 7 类种子词，这时需要利用同义词词林进行扩展评论中的相关词语，“放弃”这个词语利用同义词扩展可以得到词语集合为：

> 放弃 {甩掉，放任，遗弃，放手，舍弃，吐弃，丢弃，抛却，屏弃，抛弃，罢休，甩手，松手，停止，牺牲，摒弃，摈弃，唾弃，废弃，放胆，撒手}

这时利用同义词集合中的词语进行词语情感分类，经过计算可以将“太苦”归为“哀”。这样就通过同义词词林解决了数据稀疏性的问题，进而提高判断的准确率。

从结果中还可以看出种子词的数量对于本文改进的 SO-PMI 方法还是存在一定的影响，例如“好”、“恶”、“哀”类的种子词数量多，它的准确率、召回率、F1 值相对较高。虽然也会受到种子词数量的影响，但是采用本文提出的方法判别时，整体上比原有的 SO-PMI 方法效果好。

5.2 分类实验

建立中文图书评论情感词典的目的是为了使用该词典进行中文图书评论的情感分析。本研究采用对比实验的方法验证构建的中文图书评论情感词典的有效性，利用采集到的 100 本图书的图书评论作为语料库。对这 5 000 条图书评论分别进行分词，提取情感词。采用 4.4 节的方法进行情感类别判断，从而实现中文图书评论的情感分类。人工标注的 5 000 条图书评论的 7 类情感的分类情况和利用本文构建的中文图书评论情感词典的分类情况如表 4 所示。本组实验同样也采用准确率(P)、召回率(R)、F1 值(F1)三个评估指标评估分类方法的性能，经计算结果如表 5 所示。

表 4 5 000 条图书评论情感分类统计

情感分类	乐	好	怒	哀	惧	恶	惊
人工判别分类	668	1 396	469	561	532	1 127	247
中文图书评论词典分类	463	1 544	451	529	513	1 348	152
词典分类正确情况	437	1 384	406	489	476	1 113	138

表 5 中文图书评论情感词典情感分类效果性能评估

指标 \ 分类	乐	好	怒	哀	惧	恶	惊	总体
P	0.94	0.89	0.90	0.92	0.93	0.83	0.91	0.90
R	0.65	0.99	0.87	0.87	0.89	0.99	0.56	0.83
F1	0.77	0.94	0.88	0.89	0.91	0.90	0.69	0.85

从表 4、表 5 可以看出，采用本文构建的中文图书评论情感词典进行图书评论的情感分类的平均准确率 0.90，平均召回率为 0.83，F1 的均值为 0.85。所以能够得出使用本文构建的中文图书评论情感词典进行图书评论的情感分类具有较好的可行性和准确性。从结果中还可以看出情感词典中的词语数量对于使用情感词典进行情感分类同样也有一定的影响。情感词典中词语的数量越多，其召回率相对较高。同时实验过程中发现中文图书评论的短文本的分词和情感特征词的提取也影响情感分类的结果。

6 结 语

本文提出一种面向中文图书评论领域的情感词典构建方法，将中文图书评论的用户情感分为 7 类，提出一种改进的 SO-PMI 算法，判别中文图书评论领域情感词的情感类别，得到中文图书评论的情感词典。通过对比实验验证，本文提出的构建方法具有较好的准确性和可靠性。这种情感词典的构建方法同样也可以推广应用于其他领域情感词典的构建。

同时本研究存在一定的不足：中文图书评论短文本的分词和词频统计存在一定的误差，用户图书评论中的大量副词和连词也影响图书评论情感类别的判断；另外实验发现种子词的数量选择和语料库的规模对于情感词的情感归类也有一定的影响，如何合理地选择种子词的数量，进一步扩大语料库的规模，是后续研究和探讨的重点。

参考文献：

[1] 图书评论[EB/OL]. [2015-03-03]. http://baike.baidu.com/view/978454.htm. (Book Reviews [EB/OL]. [2015-03-03]. http://baike.baidu.com/view/978454.htm.)

[2] Andreevskaia A, Bergler S. Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses [C]. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. 2006: 209-216.

[3] Turney P D. Thumbs up or Thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews [C]. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002: 417-424.

[4] Subasic P, Huettner A. Affect Analysis of Text Using Fuzzy Semantic Typing [C]. In: Proceedings of the 9th IEEE International Conference on Fuzzy Systems. IEEE, 2001.

[5] 柳位平，朱艳辉，栗春亮，等. 中文基础情感词词典构建方法研究[J]. 计算机应用，2009，29(10)：2875-2877. (Liu Weiping, Zhu Yanhui, Li Chunliang, et al. Research on Building Chinese Basic Semantic Lexicon [J]. Journal of

Computer Application, 2009, 29(10): 2875-2877.)

[6] 李钰. 微博情感词典的构建及其在微博情感分析中的应用研究[D]. 郑州: 郑州大学, 2014. (Li Yu. Microblog Emotional Dictionary Built and Application on Sentiment Analysis of Microblog [D]. Zhengzhou: Zhengzhou University, 2014.)

[7] 桂斌, 杨小平, 张中夏, 等. 基于微博表情符号的情感词典构建研究[J]. 北京理工大学学报, 2014, 34(5):537-541. (Gui Bin, Yang Xiaoping, Zhang Zhongxia, et al. Research on Building Lexicon for Sentiment Analysis Based on the Chinese Microblogging Smiley [J]. Transactions of Beijing Institute of Technology, 2014, 34(5): 537-541.)

[8] 周咏梅, 阳爱民, 杨佳能. 一种新闻评论情感词典的构建方法[J]. 计算机科学, 2014, 41(8):67-69, 80. (Zhou Yongmei, Yang Aimin, Yang Jianeng. Construction Method of Sentiment Lexicon for New Reviews [J]. Computer Science, 2014, 41(8): 67-69, 80.)

[9] 蒋盛益, 阳垚, 廖静欣. 中文音乐情感词典构建及情感分类方法研究[J]. 计算机工程与应用, 2014, 50(24):118-121, 163. (Jiang Shengyi, Yang Yao, Liao Jingxin. Research of Building Chinese Musical Emotional Lexicon and Emotional Classification [J]. Computer Engineering and Applications, 2014, 50(24): 118-121, 163.)

[10] Yang A M, Lin J H, Zhou Y M, et al. Research on Building a Chinese Sentiment Lexicon Based on SO-PMI [J]. Applied Mechanics and Materials, 2013, 263-266: 1688-1693.

[11] 余珍芝. 中文网络产品评论的情感分析关键技术研究[D]. 杭州: 杭州电子科技大学, 2011. (Yu Zhenzhi. Research on the Key Technologies of Chinese Online Product Review's Sentiment Analysis [D]. Hangzhou: Hangzhou Dianzi University, 2011.)

[12] 李明. 面向微博电影评论的情感分类研究[D]. 昆明: 云南财经大学, 2014. (Li Ming. Emotion Classification for Weibo Movie Reviews [D]. Kunming: Yunnan Finance University, 2014.)

[13] 徐琳宏, 林鸿飞, 潘宇, 等. 情感词汇本体的构造[J]. 情报学报, 2008, 27(2):180-185. (Xu Linhong, Lin Hongfei, Pan Yu, et al. Constructing the Affective Lexicon Ontology [J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 180-185.

[14] 武汉大学 ROST 虚拟学习团队. ROST CM6 [EB/OL]. [2015-03-05]. http://download.csdn.net/download/sdaqdahai/5488041. (Virtual Learning Team of Wuhan University. ROST CM6 [EB/OL]. [2015-03-05]. http://download.csdn.net/download/sdaqdahai/5488041.)

[15] HIT-CIR Tongyici Cilin (Extended) [EB/OL]. [2015-03-05]. http://www.datatang.com/data/42306/.

[16] 知网. 《知网》情感分析用词语集: Beta [EB/OL]. [2015-03-03]. http://www.keenage.com/download/sentiment.rar. (HowNet. HowNet Sentiment Analysis Using Word Set: Beta [EB/OL]. [2015-03-03]. http://www.keenage.com/download/sentiment.rar.)

[17] 中文情感极性词典 NTUSD [EB/OL]. [2015-03-08]. http://www.datatang.com/data/44317. (Chinese Emotion Words Dictionary (NTUS) [EB/OL]. [2015-03-08]. http://www.datatang.com/data/44317.)

[18] 集搜客 GooSeeker 网络爬虫[EB/OL]. [2015-09-05]. http://www.gooseeker.com/pro/product.html. (Ji Souke GooSeeker Web Spiders [EB/OL]. [2015-09-05]. http://www.gooseeker.com/pro/product.html.)

[19] 朱嫣岚, 闵锦, 周雅倩, 等. 基于 HowNet 的词汇语义倾向性计算[J]. 中文信息学报, 2006, 20(1): 14-20. (Zhu Yanlan, Min Jin, Zhou Yaqian, et al. Semantic Orientation Computing Based on HowNet [J]. Journal of Chinese Information Processing, 2006, 20(1): 14-20.)

[20] 杜锐. 面向中文微博文本的情感分类研究[D]. 长沙: 湖南工业大学, 2014. (Du Rui. Rearch on Sentiment Classification for Chinese Microblog Text [D]. Changsha: Hunan University of Technology, 2014.)

作者贡献声明:

郭顺利: 获取、分析数据, 负责实验, 起草论文;
张向先: 提出研究思路, 设计研究方案, 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, 可通过电子邮件向作者索取, E-mail: ssguoshuli@sina.com。

[1] 郭顺利, 张向先. 8500 本图书评论数据集.rar. 8500 本总计 255000 余条的用户图书评论数据.

[2] 郭顺利, 张向先. 中文图书评论词集.rar. 881个情感词语.

[3] 郭顺利, 张向先. 情感词人工分类结果.rar. 690个情感词人工分类结果.

[4] 郭顺利, 张向先. 情感词 SO-PMI 分类结果.rar. 690情感词 SO_PMI 分类结果.

[5] 郭顺利, 张向先. 情感词改进 SO_PMI 分类结果.rar. 690情感词改进 SO_PMI 分类结果.

[6] 郭顺利, 张向先. 100本图书语料.rar. 100本图书5000条评论分类

语料.
[7] 郭顺利, 张向先. 100本图书评论人工分类结果.rar. 100本图书5000条评论人工分类结果.
[8] 郭顺利, 张向先. 100本图书评论 SO-PMI 分类结果.rar. 100本图书5000条评论 SO_PMI 分类结果.
[9] 郭顺利, 张向先. 100本图书改进 SO_PMI 分类结果.rar. 100本图书5000条评论改进 SO_PMI 分类结果.

收稿日期: 2015-09-11
收修改稿日期: 2015-10-21

Building Sentiment Analysis Dictionary for Chinese Book Reviews

Guo Shunli　Zhang Xiangxian
(School of Management, Jilin University, Changchun 130022, China)

Abstract: [**Objective**] This study aims to build a sentiment analysis dictionary for the Chinese book reviews. [**Methods**] We first divided the user's sentiments into seven categories, which were used to create the Chinese book review emotional word list. Then, chose seed terms from that list with the help of a basic sentiment analysis lexicon. Finally, used the improved SO-PMI algorithm and synonym expansion method to classify target terms from the real book reviews. [**Results**] With the help of this new book review sentiment analysis dictionary, the average precision, recall and F1 rates were 0.90, 0.83 and 0.85 respectively. [**Limitations**] The test corpus is relatively small, which might influence our results. [**Conclusions**] The proposed method was an effective and reliable way to conduct sentiment analysis for the Chinese book reviews.

Keywords: Chinese book reviews　Sentiment analysis dictionary　Seed word　Sentiment classification　SO-PMI